Clustering with Genotype and Phenotype Data to Identify Subtypes of Autism

Project Category: Life Science

Rocky Aikens (raikens)
Brianna Kozemzak (kozemzak)

December 16, 2017

1 Introduction and Motivation

Austism Spectrum Disorder (ASD) is a neurodevelopmental disorder that effects 1 in 68 individuals in the United States [1]. However, the study, diagnosis, and treatment of autism has been difficult for a variety of reasons. First, the etiology of the disease remains broadly unknown: Autism is known to be driven by a combination of genetic and environmental causes, yet few specific genetic or environmental risk factors have been identified[2]. Second, ASD can manifest over a broad spectrum of symptoms, from great intellectual and communication disability to near-normal 'high-functioning' forms. As a result, it is often asked whether ASD is in fact composed of some number of Autism 'sub-types' that are best diagnosed, studied, and treated in different ways.

Through the iHart consortium, we have access to one of the richest ASD data sets collected, containing both whole genome sequencing data and in-depth behavioral phenotype data. We aim to leverage this data set to turn machine-learning approaches toward major problems in autism research: genetic etiology, diagnosis, and sub-classification. To this end, we will address the following aims:

1. **Building an Autism Genetic Risk Score Predictor** Autism is estimated to be approximately 50% determined by genetics and 50% by environment. For these reasons, it should be possible to train a binary classifier which predicts Autism diagnosis with better-than-random performance. Such a classifier may someday be useful for identifying children at high genetic risk for autism, and analyzing the features which are most influential for building such a classifier may give us an idea of which genetic variants or gene-gene interactions are most predictive of autism diagnosis. However, because genetic information only accounts for a fraction of the factors that determine ASD development, training even an imperfect classifier is exceedingly difficult. Previous attempts at this task have yielded a maximum area under receiver operating characteristic (AU-ROC) of 0.54, a performance score only mildly better than random chance [3]. To improve upon this score, we trained a set of classification models to achieve a higher AU-ROC performance than the current state-of-the-art.
2. **Clustering Autism Sub-types** Previous research out of Stanford University [4] used Generalized Low Rank Models to cluster individuals (ASD and non-ASD) based on their behavioral profile from Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview Revised (ADI-R) examinations. However, the task of convincingly validating unsupervised learning results such as these is a common problem machine learning, since it is often unclear what defines a "good" clustering that truly reflects the underlying data-generating process (in this case, autism disease biology). Now that these clusters have been generated, substantial validation work remains to determine whether these clustering results reflect true ASD sub-types. To this end, we set about developing a pipeline to internally and externally validate these clusters, and select which of the published cluster assignments is most likely to reflect a true set of disease sub-types.

2 Building an Autism Genetic Risk Score Predictor

2.1 Data and Methods

Dataset and Feature Representation: The iHart data set includes >3,000 whole genome sequences of ASD and non-ASD individuals. However, representing the genome in a reasonable feature space is a widely considered and challenging problem. As part of a previous project, Kelley Paskov and others have developed a pipeline to identify genetic variants in these individuals which are predicted to cause loss of gene function (lof). For each individual and each likely-lof gene a 0/1 score was given: a subject receives a score of '1' for a genetic variant locus if they have two copies of the likely-lof form and 0 copies of the common form of the gene. Otherwise, the subject receives a score of 0. This allows us to represent each subject's genome as a binary vector, where each feature is a likely-lof variant locus and each person has a score of 0 or 1 for each variant. In total, this gives us >1,000 sparse binary features describing each individual in our data set for our binary classifier.

Model Development: We built a logistic regression classifier and a gradient boosted tree classifier using the `scikit learn` implementation for python 2. Twenty percent of the data was first set aside as a hold-out test set (598 individuals), and the remaining 80% (2,397 individuals) was used to develop and tune classification models using 7-fold cross-validation. Model hyperparameters (l1 regularization for logistic regression, number of estimators and maximum tree depth for gradient boosted trees) were tuned to maximize the average AU-ROC across all cross-validation folds. When the final models were selected, we retrained these on the entire training set of 2,397 individuals and evaluated their performance on the test set. For each model, test-set f1 score was calculated at whichever decision-threshold maximized f1 score over the training set.

One property of the iHart data set is that all of the recruited members are sibling pairs or trios in which at least one sibling is autistic and one is neurotypical. To account for this non-independence of genetic samples, we separated individuals into train and test sets and divided cross validation folds so that all siblings in a given family were always in the same set. This keeps our model from learning family genetic features which are uninformative. Another difficulty is that our data set contains a 2:1 ratio of cases versus controls. Ultimately, we found that oversampling our control subjects generated models which generalized more effectively rather than blindly selecting the majority class. For these reasons, we oversampled controls within each cross validation train and test set to obtain a 1:1 balance.

2.2 Results

Logistic Regression: We first built a logistic regression classifier, since this is a common standard in bioinformatics which is easy to interpret. We chose to use l1 regularization because this tends to shrink the coefficients of uninformative features to zero, generating a model which is more reflective of what we expect based on genetics. (For thoroughness, we also tuned an l2-regularized model, but found that this did not measurably change performance.) We found that a regularization parameter of $\lambda = 8$ optimized AU-ROC under cross validation. Our final model achieved an AU-ROC of 0.565 on the test set, with an f1 score of 0.634. (**Figure 1a**).

Gradient Boosted Classifier: Next, we constructed a gradient boosted tree classifier, since tree classification models, unlike logistic regression, can capture non-linear interactions between features (*i.e.* gene-gene interactions). We tuned the hyperparameters for number of trees and maximum tree depth, however we were not able to optimize additional parameters such as minimum leaf samples due to computational and time constraints. Ultimately, we found that a classifier of 40 estimators with a maximum tree depth of 2 achieved highest cross-validation performance and ultimately produced an AU-ROC of 0.58 on the test set, with an f1 score of 0.647 (**Figure 1b**).

Figure 1: ROC curves for each classification model: (a) logistic regression, (b) gradient boosted trees (c) mixed logistic regression and gradient boosted trees

Combining Models: While both models above outperform the state-of-the-art (AU-ROC of 0.54), we wondered whether we could increase performance by combining each of the models we had built. Averaging predictions from our trained logistic regression and gradient boosted tree classifiers increased our AU-ROC to 0.603 with an f1 score of 0.602 (**Figure 2c**).

2.3 Conclusions and Future Directions

Since Autism is only partially determined by genetic factors, it is impossible to perfectly predict autism diagnosis solely from the genome. This is part of the reason that previous attempts at this task have performed only slightly better than random (AU-ROC 0.54). Although the AU-ROC 0.603 we report here may seem small, it is in fact far above the current state of the art and perhaps the first suggestion that machine learning approaches may generate a useful genetic risk predictor for autism.

While it is impossible for such a predictor to perform effectively enough to replace a clinical ASD diagnosis, such a genetic risk score can provide information about an individual's genetic predisposition towards autism, and may act as a genetic 'prior' or 'prediction' in advance of diagnosis. Moreover, the features which are most predictive in a genetic autism classifier can give us an idea of which genetic variants convey the greatest risk of autism.

In the future, we would like to invest greater time and computational resources into optimizing ensemble-based methods for autism prediction (for example, by further hyperparameter optimization of the gradient boosted tree classifier), and express additional genomic samples in our binary likely-lof feature representation so that they can be used as extra training data. Each of these steps should help us build more generalizable models which can be used to generate an Autism genetic risk score for un-diagnosed individuals. Once our most effective models are tuned and trained, we can then analyze the predictive features in these models to generate hypotheses regarding which genetic elements or gene-gene relationships are most fundamental to deciding autism risk.

3 Clustering Autism Subtypes

3.1 Data and Methods

Dataset: Through iHart, we have access to Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview Revised (ADI-R) results for 13,434 individuals. These examinations are typically administered by medical professionals to individuals with suspected ASD to capture their behavioral profile, so even the "controls" in this dataset are unlikely to be completely neurotypical. However, mid-way through this project, we were able to incorporate data for an additional 59 individuals, representing a small sample of carefully-selected neurotypical controls. In addition to this phenotypic data, we have 29 labels for the individuals, including their data source, diagnosis, demographic information, computed ADOS and ADI-R categorical scores, and pertinent medical history.

Prior Work: As part of a previous project, Kelley Paskov and others performed data pre-processing and generated clustering results for the dataset described above. They reduced the feature space by removing highly correlated features and by only using features with binary or ordinal values, yielding a total of 123 numerical features. Next, they simultaneously imputed missing data values and generated kmeans and soft kmeans clusters for $k = 2, 3, ..., 6$ using a Generalized Low Rank Model with logistic loss.

Methods: The validation of unsupervised learning methods can be challenging and the methods used often depend on the initial motivation for performing the unsupervised learning task. In this case, we are interested in determining whether the clustering results generated by Kelley Paskov and others reflect true ASD subtypes. To address this question, we analyzed three different aspects of the clusters:

1. *Features:* The feature values usually represent the extent to which a behavior is observed, where lower values are associated with the neurotypical case. After scaling the features to be in the interval $[0, 1]$, we can visualize the degree to which individuals display neurotypical or atypical behaviors within each cluster using a heatmap of feature values for each individual organized by cluster membership. For a more succinct representation, we use cluster centroids, which describe the average behavior of the individuals in a cluster.
2. *Labels:* We create pie charts showing the various label proportions between clusters and perform multiple hypergeometric tests with a Benjamini-Hochberg correction to determine if any label values are significantly enriched for certain clusters. This gives us information about characteristics of the individuals that the clusters may be capturing, including diagnosis, gender, medical history, etc.

3. *Movement:* To prevent bias in the structure of the labels from limiting our analysis, we examine the movement of individuals between clusters as k increases. This not only allows us to observe how the composition of clusters changes with the selection of a parameter k, but also enables us to determine if individuals are being assigned to clusters in a non-random, biologically-relevant way.

We apply these method to the kmeans and soft kmeans clustering results for $k = 2, 3, ..., 6$ using both the original dataset and the dataset updated with neurotypical controls. For soft kmeans, we assign individuals to a single cluster based on their maximum partial membership.

3.2 Results

K-Means: *Features:* For both the heatmap of individuals sorted by cluster and the heatmap of cluster centroids, there are no obvious patterns in the features values until $k = 4$, where a cluster forms with lower feature values corresponding to more neurotypical behavior. As k increases, the neurotypical cluster remains present, but there are no observable differences between the feature values of the other clusters (see Figure 2). *Labels:* The pattern above is also observed in the label pie charts, where the cluster label proportions are almost indistinguishable until $k = 4$. Even the added neurotypical controls, which should have significantly different feature values, do not all cluster together until $k = 4$. *Movement:* When $k = 2$ and 3, the proportion of individuals moving to each of the clusters for $k + 1$ is equal across all clusters, leading us to believe that the individuals are being assigned to clusters arbitrarily (see Figure 3). When $k = 4$ and 5, the proportion of individuals moving to each of the clusters for $k + 1$ is equal for one or more (but not all) clusters, leading us to believe that the divisions between some clusters are not biologically meaningful (see Figure 4).

Soft K-Means: *Features:* For every $k = 2, 3, , 6$, the behavioral feature values between clusters appear to form a gradient ranging from neurotypical to 'low-functioning' ASD, where the clusters are separated by severity of ASD (see Figure 5). *Labels:* The label pie charts show biologically meaningful divisions between clusters for all values of k. While not every cluster was significantly different for every label, at least one label was significantly different for each cluster and each k. We observe some differences in the clustering results after the addition of the neurotypical controls. Take, for example, the case when $k = 3$ (see Figure 6). Although these differences exist, the broader patterns of differing autism severity are evident in the communication, social interaction, and restricted repetitive behavior label proportions. Additionally, the neurotypical controls are always together in a single cluster, lending validity to the clusters ability to

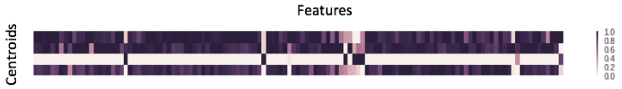

Figure 2: Heatmap of features versus centroids for kmeans with $k = 4$

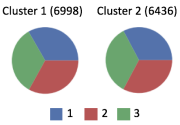

Figure 3: Movement from $k = 2$ to $k = 3$ for kmeans

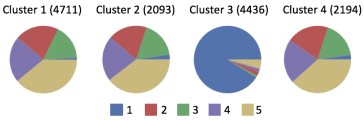

Figure 4: Movement from $k = 4$ to $k = 5$ for kmeans

Figure 5: Heatmap of features versus centroids for soft kmeans with $k = 3$

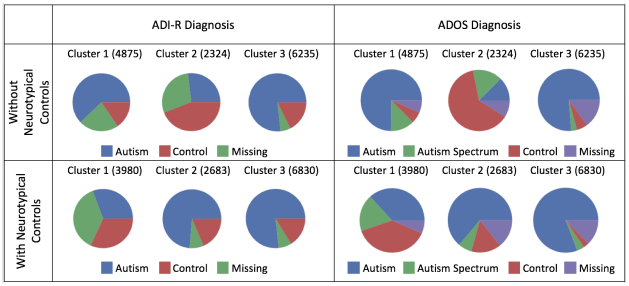

Figure 6: Comparison of ADOS and ADI-R diagnosis for original dataset and dataset with added neurotypical controls for soft kmeans with $k = 3$

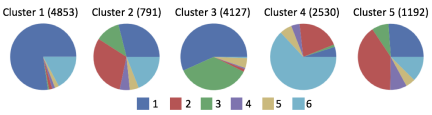

Figure 7: Movement from $k = 5$ to $k = 6$ for soft kmeans

capture diagnosis. *Movement:* The proportion of individuals moving to each of the clusters for $k+1$ is different across all clusters for all $k = 2, 3, , 5$. This pattern suggests that the movement between clusters is related to some underlying phenotypic characteristic of the cluster (see Figure 7).

3.3 Conclusions and Future Directions

Taking all results into account, we conclude that the clusters formed using kmeans are not biologically relevant. An analysis of the composition of clusters based on the dataset that the individuals came from and the number of missing feature values did not seem to indicate any batch effects causing these undesirable results. However, we feel comfortable concluding that subtypes of autism can be identified from the soft clustering results and that there may be more than 6 valid subtypes, since we never observed the formation of 'extra clusters causing divisions between two or more clusters to not be biologically meaningful.

In the future, we would like to perform soft kmeans clustering for values of $k > 6$, since we never observed the formation of 'extra clusters in the previously-generated results. Additionally, we would like to implement methods for working directly with the soft clustering results rather than having to impose crisp memberships. This would include creating pie charts where counts are weighted by cluster membership and measuring cluster label enrichment using the Wald test statistic for soft clustering [5]. Lastly, we would like to generate volcano plots for our calculated label enrichment p-values to help identify which tests were most significant, because there were too many significant tests to analyze all of them manually.

4 Final Discussion

Herein, we discuss two applications of machine learning to classical ASD research problems. First, we build an Autism genetic risk predictor, which outperforms current state-of-the-art (AU-ROC of 0.60 compared to 0.54). This result indicates that a reduced representation of the genome can indeed predict ASD with better than random chance. Further work collecting additional training data and optimizing ensemble-based classification models can build an effective genetic ASD predictor which can be used to estimate a child's genetic risk of autism and identify genetic factors most predictive of ASD. Second, we develop a pipeline to externally and internally validate previously generated ASD cluster assignments of $> 13,000$ individuals. Based on several metrics, we find that a soft clustering into $k = 3$ groups appears to best separate individuals into ASD subtypes which appear to be potentially biologically meaningful. Further work building and evaluating smarter and smarter cluster assignments may someday help us develop a new subclassification of autism, so that these groups of invididuals can be researched, diagnosed, and treated more effectively based on their specific disease profile. These promising early results in both aims suggest that further work collecting rich ASD and control data sets for machine learning approaches will help us address outstanding issues in the study and treatment of autism.

Contributions

Rachael Aikens has primarily headed work on Aim 1: Building an Autism Genetic Risk Score Predictor. Brianna Kozemzak has primarily headed work on Aim 2: Clustering Autism Subtypes. Each student has played an equal part in drafting the final report, milestone, poster, and project proposal. Professor Dennis Wall helped advise the both authors on approaching these problems and developing research aims. Kelley Paskov and Nate Stockholm (two Wall lab members outside of CS 229) had previously built the original imputed phenotype data, previous clustering results, and binary representations of genetic information as parts of prior research projects.

References

[1] Disabilities Monitoring Network Surveillance Year Developmental, 2010 Principal Investigators, et al. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, 63(2):1, 2014.

[2] Kristen Lyall, Lisa Croen, Julie Daniels, M Daniele Fallin, Christine Ladd-Acosta, Brian K Lee, Bo Y Park, Nathaniel W Snyder, Diana Schendel, Heather Volk, et al. The changing epidemiology of autism spectrum disorders. *Annual review of public health*, 38:81–102, 2017.

[3] Wall Lab Group. Augur: Predicting autism phenotype from genotype. unpublished.

[4] Kelley M. Paskov and Dennis P. Wall. Low rank methods for phenotype imputation and clustering in autism spectrum disorder. unpublished.

[5] R.D. Phillips, M.S. Hossain, L.T. Watson, R.H. Wynne, and Naren Ramakrishnan. Enrichment procedures for soft clusters: A statistical test and its applications. *Computer Modeling in Engineering and Sciences*, 97:175–197, 2014.